Blood-derived mitochondrial DNA copy number is associated with gene expression across multiple tissues and is predictive for incident neurodegenerative disease

Stephanie Y. Yang,[1] Christina A. Castellani,[1] Ryan J. Longchamps,[1] Vamsee K. Pillalamarri,[1] Brian O'Rourke,[2] Eliseo Guallar,[3] and Dan E. Arking[1,2]

[1]*McKusick-Nathans Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA;* [2]*Division of Cardiology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA;* [3]*Departments of Epidemiology and Medicine, and Welch Center for Prevention, Epidemiology, and Clinical Research, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland 21205, USA*

Mitochondrial DNA copy number (mtDNA-CN) is a proxy for mitochondrial function and is associated with aging-related diseases. However, it is unclear how mtDNA-CN measured in blood can reflect diseases that primarily manifest in other tissues. Using the Genotype-Tissue Expression Project, we interrogated relationships between mtDNA-CN measured in whole blood and gene expression from whole blood and 47 additional tissues in 419 individuals. mtDNA-CN was significantly associated with expression of 700 genes in whole blood, including nuclear genes required for mtDNA replication. Significant enrichment was observed for splicing and ubiquitin-mediated proteolysis pathways, as well as target genes for the mitochondrial transcription factor NRF1. In nonblood tissues, there were more significantly associated genes than expected in 30 tissues, suggesting that global gene expression in those tissues is correlated with blood-derived mtDNA-CN. Neurodegenerative disease pathways were significantly associated in multiple tissues, and in an independent data set, the UK Biobank, we observed that higher mtDNA-CN was significantly associated with lower rates of both prevalent (OR = 0.89, CI = 0.83; 0.96) and incident neurodegenerative disease (HR = 0.95, 95% CI = 0.91;0.98). The observation that mtDNA-CN measured in blood is associated with gene expression in other tissues suggests that blood-derived mtDNA-CN can reflect metabolic health across multiple tissues. Identification of key pathways including splicing, RNA binding, and catalysis reinforces the importance of mitochondria in maintaining cellular homeostasis. Finally, validation of the role of mtDNA CN in neurodegenerative disease in a large independent cohort study solidifies the link between blood-derived mtDNA-CN, altered gene expression in multiple tissues, and aging-related disease.

[Supplemental material is available for this article.]

Mitochondria perform multiple essential metabolic functions including energy production, lipid metabolism, and signaling for apoptosis. Mitochondria possess circular genomes (mtDNA) that are distinct from the nuclear genome. Although cells typically only possess two copies of the nuclear genome, they contain 100s to 1000s of mitochondria, and each individual mitochondrion can hold 2–10 copies of mtDNA, resulting in wide variation in mtDNA copy number (mtDNA-CN) (Wai et al. 2010). The amount of mtDNA-CN also varies widely across cell types, with higher energy demand cell types typically possessing higher levels of mtDNA-CN (Chabi et al. 2003; Miller et al. 2003; Clay Montier et al. 2009; Kelly et al. 2012). Due to the importance of mitochondria in metabolism and energy production, mitochondrial dysfunction plays a role in the etiology of many human diseases (Herst et al. 2017). mtDNA-CN has been shown to be a proxy for mitochondrial function and is consequently an attractive biomarker due to its ease of measurement (Malik and Czajka 2013; Castellani et al. 2020b). Indeed, low levels of mtDNA-CN in peripheral blood have been associated with an increased risk for a number of chronic aging-related diseases including frailty, kidney disease, cardiovascular disease, heart failure, and overall mortality (Ashar et al. 2015, 2017; Huang et al. 2016; Tin et al. 2016).

Crosstalk between the mitochondrial and nuclear genomes is essential for maintaining cellular homeostasis. Many essential mitochondrial proteins are encoded by the nuclear genome, and expression of these nuclear genes must be modified to match mitochondrial activity. Likewise, mitochondrial activity must respond to cellular energy demands. Polymorphisms in the nuclear genome have been associated with changes in mitochondrial gene expression, and mitochondrial genome variation has been associated with changes in nuclear gene expression, suggesting interplay between the two genomes (Lee et al. 2017b; Ali et al. 2019).

In cancer cells, mtDNA-CN alters gene expression through modifying DNA methylation (Reznik et al. 2016; Sun and St John 2018). Recent work from our lab has shown that mtDNA-

Corresponding author: arking@jhmi.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.269381.120.

CN is also associated with nuclear DNA methylation in noncancer settings (Castellani et al. 2020a). Given that DNA methylation can modify gene expression, the current study seeks to explore the potential association between blood-derived mtDNA-CN and gene expression. Past work has shown that mtDNA-CN is associated with gene expression of nuclear-encoded genes in lymphoblast cell lines, but this may not reflect biological processes occurring in other tissues, especially after an extended culturing period (Gibbons et al. 2014). Therefore, we leveraged data from the Genotype-Tissue Expression Project (GTEx), a cross-sectional study with gene expression data from multiple nondiseased postmortem tissues, to examine associations between mtDNA-CN and expression of both nuclear and mitochondrially encoded genes (The GTEx Consortium 2013). This study aimed to evaluate associations between blood-derived mtDNA-CN and gene expression across multiple tissues and to follow up on a novel association between neurodegenerative disease and blood-derived mtDNA-CN.

Results

Determination and validation of mtDNA-CN metric

mtDNA-CN estimates were generated from whole genome sequences performed on DNA derived from whole blood using the ratio of mitochondrial reads to total aligned reads. As mtDNA-CN is known to be affected by cell type composition, cell counts for samples with available RNA-sequencing data were deconvoluted using gene expression measured in whole blood (Aran et al. 2017; Zhang et al. 2017). We identified a batch effect that resulted in significantly altered mtDNA-CN for individuals sequenced prior to January 2013. Therefore, only individuals sequenced after January 2013 were retained for analysis (Supplemental Fig. S1). After quality control, outlier filtering, and normalization of the RNA-sequencing data, 419 individuals remained for analyses (see Methods).

To validate mtDNA-CN measurements in the filtered GTEx data, we determined the association between mtDNA-CN and known correlated measures, including age, sex, and neutrophil count (Mengel-From et al. 2014; Zhang et al. 2017; Moore et al. 2018). We observed a significant association with neutrophil count ($P = 8.4 \times 10^{-5}$), with higher neutrophil count associated with lower mtDNA-CN. Although not statistically significant, effect size estimates between mtDNA-CN and age ($P = 0.18$) and sex ($P = 0.14$) were also in the expected direction, with older individuals and males having lower mtDNA-CN (Supplemental Fig. S2). Effect size estimates for age and neutrophils were also consistent with prior literature (Supplemental Table S1; Longchamps et al. 2020). Based on variance explained from previous studies, the current study was only powered to detect a significant effect for neutrophil count. For all downstream analyses, mtDNA-CN was defined as the standardized residual from a linear regression model adjusted for age, sex, cell counts estimated from RNA-seq deconvolution, ischemic time, and cohort (see Methods).

Association of mtDNA-CN derived from whole blood with gene expression in blood

A priori, we expect that mitochondrially encoded gene expression would be positively correlated with mtDNA-CN. Likewise, multiple nuclear-encoded genes are involved in the regulation of mtDNA replication, and thus, expression levels of these genes are expected to be correlated with mtDNA-CN (Garcia et al. 2017; Rusecka et al. 2018). We therefore evaluated the associations between mtDNA-CN and expression of these two classes of genes, correcting for cohort, sample ischemic time, genotyping PCs, age, race, and surrogate variables derived from RNA-sequencing data to capture known and hidden confounders (Supplemental Fig. S3; Leek and Storey 2007).

To minimize the potential impact of outliers, we performed an inverse normal transformation on both the mtDNA-CN metric and the gene expression values. To evaluate the association between mtDNA-CN and mitochondrial RNA (mtRNA) levels, we used the median gene expression value calculated from scaled expression values across 36 mtDNA-encoded genes that passed expression thresholds (see Methods).

We observed a highly significant association between mtDNA-CN and overall mtRNA expression ($P = 9.10 \times 10^{-9}$) (Table 1), with 33 out of 36 individual mtDNA-encoded genes nominally significant ($P < 0.05$) (Supplemental Fig. S4).

In addition to genes coding directly for mtDNA replication machinery, genes involved in mtDNA transcription and nucleotide metabolism are also required for mtDNA replication. The mtDNA transcription machinery provides the RNA primers used in mtDNA replication, and nucleotides are needed to synthesize new mtDNA molecules. Of the 17 mtDNA major replication genes tested (Rusecka et al. 2018), all were positively associated with mtDNA-CN, as would be expected based on gene function; eight of them were nominally significant ($P < 0.05$), and four were significant after Bonferroni correction ($P < 2.94 \times 10^{-3}$) for multiple testing (Table 1).

To identify additional genes and pathways associated with mtDNA-CN, we performed a transcriptome-wide analysis. There was an overall inflation of test statistics, which we quantified using

Table 1. Blood-derived mtDNA-CN is positively associated with gene expression for mitochondrially encoded genes and nuclear-encoded genes required for mtDNA replication

Gene	Effect size estimate	Standard error	*P*-value
Scaled mtRNA median	0.15	0.03	9.10×10^{-9}
mtDNA replication machinery			
POLG	0.02	0.01	0.025
POLG2	0.06	0.02	4.08×10^{-4}
TWNK	0.03	0.02	0.11
SSBP1	0.06	0.01	1.38×10^{-4}
PRIMPOL	0.04	0.02	0.020
DNA2	0.05	0.02	0.010
MGME1	0.04	0.02	0.048
RNASEH1	0.06	0.02	2.51×10^{-4}
mtDNA transcription machinery			
TFAM	0.06	0.01	1.83×10^{-4}
TEFM	0.03	0.02	0.05
TFB2M	0.03	0.02	0.028
POLRMT	0.01	0.01	0.19
Nucleotide metabolism genes			
TK2	0.02	0.02	0.11
DGUOK	0.06	0.02	6.39×10^{-4}
RRM2B	0.04	0.01	0.006
TYMP	0.02	0.02	0.29
SLC25A4	0.08	0.03	0.003

Effect size estimates represent the change in gene expression, in standard deviation units, associated with a 1-standard-deviation increase in blood-derived mtDNA-CN. Mitochondrially encoded genes are represented as the median of the scaled mtRNA expression of the 36 genes with detectable expression. Genes required for mtDNA replication were obtained from Rusecka et al. (2018).

the genomic inflation factor (lambda = 4.71) (Devlin and Roeder 1999). Two-stage permutation testing with 1000 permutations demonstrated no inflation in null data sets, suggesting that this inflation represents a true global association between blood-derived mtDNA-CN and gene expression (Supplemental Fig. S5).

When stratified by gene functional categories (Harrow et al. 2012), all categories showed elevated test statistics, but protein-coding genes were the most enriched (lambda = 7.44) (Fig. 1).

Gene expression levels of most of the nominally significant genes were positively correlated with mtDNA-CN (7769 genes with positive t-values vs. 285 genes with negative t-values). Although much of this positive skewing is likely due to correlated gene expression, permuted data sets demonstrate that this positive shift is significant ($P < 0.001$) (Supplemental Fig. S6), perhaps reflecting a more active transcriptional state associated with higher mtDNA-CN. Whereas 698 of the significantly associated genes were positively associated, only two negatively associated genes passed the permutation cutoff ($P < 2.38 \times 10^{-6}$), *CAMP* ($P = 1.58 \times 10^{-8}$), and *PGLYRP1* ($P = 1.78 \times 10^{-7}$), both of which are involved in innate immunity.

Gene set enrichment analysis uncovers gene regulatory networks in whole blood

To identify specific molecular pathways, transcription factors, and gene ontologies associated with mtDNA-CN in whole blood, we performed gene set enrichment analyses (Irizarry et al. 2009) using gene sets obtained from the Molecular Signatures Database (MSigDB) (Ashburner et al. 2000; Xie et al. 2005; Liberzon et al. 2011; Meng et al. 2019; The Gene Ontology Consortium 2019). Previous studies have shown that cross-mappability can lead to false pseudogene positives in eQTL association studies (Saha and Battle 2019); we therefore excluded pseudogenes from subsequent analyses. Significantly associated KEGG pathways included "Spliceosome" ($P = 1.03 \times 10^{-8}$) and "Ubiquitin-mediated proteolysis" ($P = 2.4 \times 10^{-10}$) (Table 2).

A number of transcription factor target sequences were also significantly enriched, including those for ELK1 ($P = 8.58 \times 10^{-66}$), NRF1 ($P = 1.76 \times 10^{-35}$), GABPB ($P = 3.54 \times 10^{-21}$), YY1 ($P = 3.14 \times 10^{-19}$), and E4F1 ($P = 3.98 \times 10^{-15}$). All of these transcription factors regulate genes that play a role in mitochondrial function (Barrett et al. 2006; Blesa et al. 2008; Yang et al. 2014; Rodier et al. 2015; Chen et al. 2019). Gene expression levels of these transcription factors were all positively correlated with mtDNA-CN, with five out of six nominally significant, and three remaining significant after Bonferroni correction ($P < 8.33 \times 10^{-3}$) (Table 3).

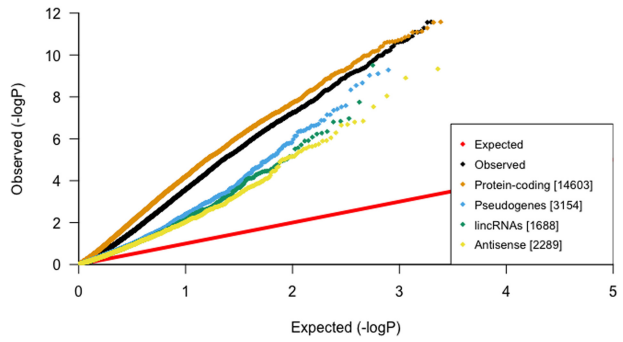

Figure 1. Global inflation of test statistics from linear regressions between blood-derived mtDNA-CN and gene expression in blood. After stratification by gene category, protein-coding genes have the most inflation, suggesting that mtDNA-CN is strongly associated with genes that code for proteins.

Table 2. Top five genes that were most significantly associated with mtDNA-CN within the "Spliceosome" and "Ubiquitin-mediated proteolysis" KEGG pathways

Gene	Effect size estimate	Standard error	*P*-value
Spliceosome genes			
TRA2A	0.11	0.01	2.99×10^{-14}
LSM6	0.11	0.02	3.75×10^{-10}
HNRNPA1L2	0.12	0.02	2.07×10^{-8}
SRSF8	0.10	0.02	2.24×10^{-7}
NCBP2	0.06	0.01	6.60×10^{-7}
Ubiquitin-mediated proteolysis genes			
UBE2B	0.12	0.02	1.20×10^{-13}
ELOC	0.08	0.01	2.02×10^{-8}
UBE2I	0.09	0.02	6.26×10^{-8}
CUL1	0.07	0.01	9.73×10^{-8}
UBE2K	0.07	0.01	1.32×10^{-7}

Effect size estimates represent the change in gene expression, in standard deviation units, associated with a 1- standard-deviation increase in blood-derived mtDNA-CN.

Many mitochondrially related cellular component Gene Ontology (GO) terms were significant, including "Mitochondrion" ($P = 7.77 \times 10^{-23}$), "Mitochondrial part" ($P = 2.79 \times 10^{-15}$), and "Mitochondrion organization" ($P = 2.87 \times 10^{-14}$) (Fig. 2; Supek et al. 2011).

Additional significantly associated GO terms included "ubiquitin ligase complex" ($P = 6.6 \times 10^{-18}$) and "spliceosomal complex" ($P = 4.46 \times 10^{-14}$), supporting the KEGG pathway findings. Genes with substantial evidence of mitochondrial localization, determined through integration of several genome-scale data sets, were obtained from MitoCarta2.0 and demonstrated significant enrichment ($P = 8.22 \times 10^{-21}$) (Calvo et al. 2016).

Cross-tissue analysis reveals associations between gene expression in multiple tissues and blood-derived mtDNA-CN

mtDNA-CN measured in blood has been associated with a number of aging-related diseases, including chronic kidney disease, heart failure, and diabetes (Huang et al. 2016; Tin et al. 2016; Al-Kafaji et al. 2018). Given that these diseases primarily manifest in non-blood tissues, we evaluated associations between blood-derived mtDNA-CN and gene expression measured from 47 additional tissues that had greater than 50 samples after filtering.

Though blood-derived mtDNA-CN appears to be associated with gene expression in other tissues, we did not observe a significant association between blood-derived mtDNA-CN and scaled mtRNA gene expression in any tissue other than blood, and only two out of 47 tested tissues had nominally significant associations between tissue-specific scaled mtRNA expression and blood-derived mtDNA-CN (uterus [$P = 0.004$], left ventricle of the heart [$P = 0.017$]) (Supplemental Table S2). However, mtRNA expression for 35/47 nonblood tissues was positively associated with blood mtDNA-CN, which is more than what would be expected by chance ($P < 0.001$). This suggests that, although our study may be underpowered to detect a significant association in individual tissues due to small sample sizes, mtDNA-CN measured in blood is broadly correlated with mtDNA-CN in other tissues. As expected, mtRNA expression varies widely across tissues, with brain tissues having notably more expression than other tissues (Supplemental Fig. S7).

Table 3. Gene expression for transcription factors whose targets are enriched for association with blood-derived mtDNA-CN is nearly all nominally significantly associated with blood-derived mtDNA-CN

Gene	Effect size estimate (gene expression)	Standard error (gene expression)	*P*-value (gene expression)	*P*-value (enriched target sequences)
NRF1	0.03	0.02	0.07	1.76×10^{-35}
YY1	0.07	0.01	1.78×10^{-6}	3.14×10^{-19}
GABPB2	0.09	0.02	1.51×10^{-9}	3.54×10^{-21}
GABPB1	0.03	0.01	0.048	3.54×10^{-21}
E4F1	0.05	0.01	2.01×10^{-4}	3.98×10^{-15}
ELK1	0.04	0.02	0.021	8.58×10^{-66}

Effect size estimates, standard errors, and *P*-values from a linear regression between transcription factor gene expression and blood-derived mtDNA-CN. Transcription factors shown are those whose targets were significantly enriched for association with blood-derived mtDNA-CN.

We calculated genomic inflation factors for each tissue to quantify test statistic inflation. Genomic inflation factors were elevated across multiple nonblood tissues, suggesting that blood-derived mtDNA-CN was broadly associated with gene expression in other tissues (Fig. 3).

To determine true signal from noise, we performed 1000 two-stage permutations for each tissue and obtained a genomic inflation factor lambda cutoff of >1.20 representing a significant elevation of lambda (study-wide $P < 0.05$). Using this cutoff, we identified 30 nonblood tissues with a global inflation of test statistics (Supplemental Table S3). Other than blood, the most strongly enriched tissue was the putamen region of the brain, with a lambda of 3.27. Principal components analysis revealed that the putamen region of the brain was not significantly different from other brain regions, and we are uncertain what is biologically causing the strong enrichment (Supplemental Fig. S8). We note that the two cell lines, EBV transformed lymphocytes (lambda = 0.84) and cultured fibroblasts (lambda = 0.84), showed no global inflation of test statistics, suggesting that blood-derived mtDNA-CN loses its association with gene expression after the cell culturing process.

To examine the similarity of associations of mtDNA-CN observed in blood with other tissues, we calculated Spearman's rank correlation coefficients between effect estimates for blood-derived mtDNA-CN on blood gene expression (β_{blood}) and effect estimates for blood-derived mtDNA-CN on gene expression in other tissues (β_{tissue}). All genes that passed a permutation cutoff for significance in blood ($P = 2.38 \times 10^{-6}$, 700 genes total) were included. To distinguish tissues with correlations more extreme than baseline, we calculated reference correlations between blood and other tissues for 1000 randomly selected sets of 700 genes. Twenty-six tissues had observed values that were significantly more extreme than baseline (Supplemental Table S4), with 22 tissues showing greater correlation and four tissues showing less correlation. Of these 26 tissues, 20 were among the 30 tissues with significantly inflated lambdas.

To identify pathways associated with mtDNA-CN across multiple tissues, we performed gene set enrichment analysis in each of the 30 tissues with a significant genomic inflation factor. Multiple terms were significant in greater than one tissue (Table 4), including terms related to oxidative phosphorylation and mitochondria, suggesting that mtDNA-CN derived from blood can reflect mitochondrial function occurring in other tissues.

ELK1 transcription factor binding sites were significantly enriched in 19 of the 30 significant tissues and were also significant in whole blood, suggesting that mtDNA-CN may regulate ELK1 or vice versa. We note that gene expression for *ELK1* was nominally significantly associated ($P < 0.05$) with blood-derived mtDNA-CN in four of the 18 tissues for which ELK1 targets were significantly enriched (Supplemental Fig. S9). Effect estimates for ELK1 targets were generally consistent with the directionality of ELK1 effect estimates. For example, in blood, where *ELK1* expression is positively associated with mtDNA-CN, 747/750 (99.6%) nominally significant ELK1 target genes were positively associated. On the other hand, mtDNA-CN was negatively associated with nerve *ELK1* gene expression, and 204/306 (66.67%) nominally significant ELK1 target genes were also negatively associated. Of note, nearly all the noted transcription factors are ubiquitously expressed throughout the body, except for *ELK1*, which is not expressed in brain putamen or spinal cord (Supplemental Fig. S10).

To identify genes driving enrichment of significant pathways in multiple tissues, we performed a random effects meta-analysis for all expressed genes using effect size estimates from all 47 nonblood tissues. Genes encoding both the large and small ribosomal subunits were negatively associated with blood-derived mtDNA-CN across all tested tissues, implying an inverse relationship between ribosomal abundance and mitochondrial DNA quantity (Table 5).

Huntington's disease (HD), Parkinson's disease (PD), and Alzheimer's disease (AD) were among the most significantly associated KEGG pathways that appear in multiple tissues (Table 4). This is an intriguing finding, given the known role of mitochondria in neurodegenerative disease (Reddy 2009).

Although neurodegenerative disorders primarily manifest in nervous tissues (Wood et al. 2015), we observed significant enrichment of disease pathways in colon, pancreas, and testis tissues. When limiting our query to brain tissues, HD and PD were nominally significantly enriched in cerebellum, caudate (basal ganglia), and cortex, whereas AD was nominally significantly enriched in cerebellum and spinal cord (Supplemental Fig. S11).

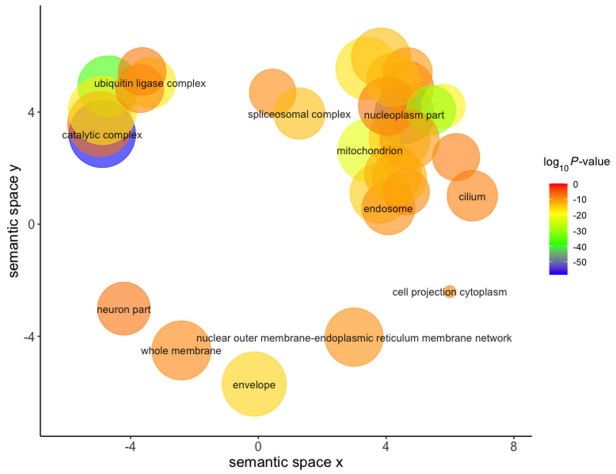

Figure 2. REVIGO visualization of GO cellular component terms significantly associated with mtDNA-CN after removal of redundant GO terms. Size of the circle represents the relative number of genes in each gene set, color represents significance. Axes represent semantic similarities between GO terms; GO terms that are more similar will cluster with one another.

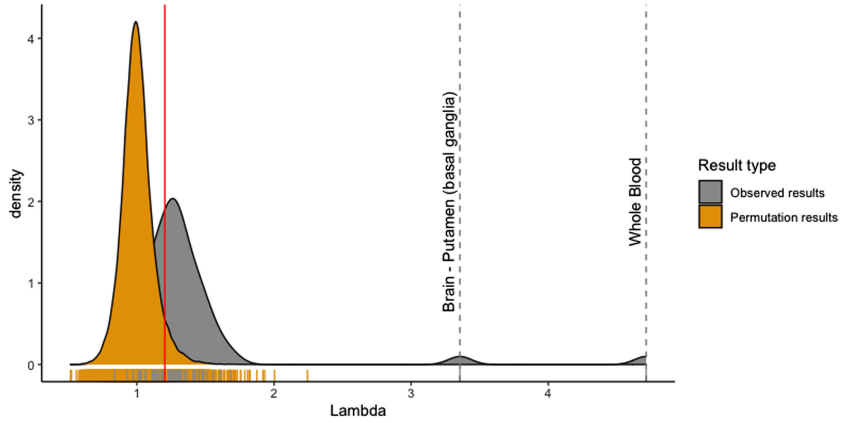

Figure 3. Observed genomic inflation factors are significantly different from permuted genomic inflation factors for certain tissues. A higher genomic inflation factor represents increased global associations between blood-derived mtDNA-CN and gene expression in a specific tissue. One thousand permuted genomic inflation factors were obtained using two-stage permutation testing. Red line represents permuted lambda cutoff of 1.20.

mtDNA-CN is associated with incident and prevalent neurodegenerative disease in the UK Biobank

To examine the association between mtDNA-CN and neurodegenerative disease risk, we used the UK Biobank (UKB) (Bycroft et al. 2018), a prospective cohort study with whole exome sequencing (WES) for ~50,000 individuals and genotyping for ~500,000 individuals. mtDNA-CN was estimated from a combination of WES data and mitochondrial SNP probe intensities from genotyping arrays (see Methods) and was significantly associated with age and sex in the expected directions (Supplemental Fig. S12). Analyses were restricted to unrelated individuals of self-identified European descent, and individuals with blood cell type count outliers were excluded, followed by adjustment of mtDNA-CN for age and sex. Associations between mtDNA-CN and prevalent and incident neurodegenerative disease were evaluated using logistic regression models and Cox proportional-hazards models, respectively. mtDNA-CN was significantly associated with prevalent Parkinson's disease and prevalent dementia (Table 6). As there were only 12 individuals with prevalent Alzheimer's disease, we did not test for an association with prevalent Alzheimer's disease. For incident disease, median follow-up time was ~10 yr. mtDNA-CN was significantly associated with incident Parkinson's disease and incident dementia (Table 6). Consistent with other aging-related diseases (Ashar et al. 2015, 2017), higher mtDNA-CN was associated with lower risk for developing incident neurodegenerative disease (Table 6). A combined analysis for all individuals with neurodegenerative disease revealed a consistent strongly significant association for both prevalent (OR = 0.89, CI = 0.83;0.96) and incident (HR = 0.95, CI = 0.91;0.98) disease.

Discussion

In this study, blood-derived mtDNA-CN was significantly associated with a host of blood-expressed genes. As expected, nearly all genes involved in mtDNA replication were significantly associated with mtDNA-CN in a positive direction. There was also a clear overall shift toward significant positive estimates, possibly indicating that increased mtDNA-CN is reflective of a more active transcriptional state. This finding is consistent with previous literature demonstrating that higher mitochondrial content is correlated with increased transcriptional activity (Guantes et al. 2015; Márquez-Jurado et al. 2018). The two negatively associated genes both play roles in innate immune function (Gombart et al. 2005; Osanai et al. 2011), suggesting that higher mtDNA-CN levels are correlated with decreased immune response. Mitochondria play a role in immune responses to pathogens in several ways; for example, mitochondrial DNA release from compromised mitochondria can trigger an intracellular antiviral response through the cGAS–STING pathway (West et al. 2015), binding of viral dsRNA to the mitochondrial antiviral signaling complex (MAVS) can trigger an interferon responses through STAT6 activation (Chen et al. 2011), and release of mitochondrial components from cells can bind to damage-associated molecular pattern (DAMP) receptors to trigger innate immune responses (Nakahira et al. 2015). These novel findings correlating expression of mtDNA-CN with specific immune response genes in tissues represent an area for further investigation.

Gene set enrichment analyses revealed pathways potentially involved in mitochondrial DNA control, including ubiquitin-mediated proteolysis and splicing. Supporting this finding, Guantes et al. demonstrated that mitochondrial content modulates alternative splicing (Guantes et al. 2015). Additionally, we found that genes expressed in whole blood that were associated with blood-derived mtDNA-CN were enriched for target sequences for the ELK1,

Table 4. Pathways, transcription factor targets, and GO terms significantly enriched in multiple tissues (excluding blood)

Pathway	Number of significant tissues
Transcription factors	
SCGGAAGY_ELK1_02	19
RCGCANGCGY_NRF1_Q6	12
GCCATNTTG_YY1_Q6	8
TGCGCANK_UNKNOWN	8
MGGAAGTG_GABP_B	7
GO terms	
GO_RNA_BINDING	16
GO_CATALYTIC_COMPLEX	14
GO_CELLULAR_MACROMOLECULE_LOCALIZATION	13
GO_INTRACELLULAR_TRANSPORT	13
GO_MACROMOLECULE_CATABOLIC_PROCESS	13
KEGG terms	
KEGG_RIBOSOME	7
KEGG_HUNTINGTONS_DISEASE	5
KEGG_OXIDATIVE_PHOSPHORYLATION	4
KEGG_PARKINSONS_DISEASE	4
KEGG_ALZHEIMERS_DISEASE	3
Mitochondrial terms	
GO_MITOCHONDRIAL_ENVELOPE	8
GO_MITOCHONDRIAL_PART	8
GO_MITOCHONDRION	8
GO_MITOCHONDRION_ORGANIZATION	4
GO_MITOCHONDRIAL_PROTEIN_COMPLEX	3

The top five terms for each category that were significantly enriched in multiple tissues are shown.

Table 5. Random-effects meta-analysis for genes driving the enrichment of pathways in multiple tissues

Gene	Meta effect size estimate	Meta standard error	Meta *P*-value
ELK1 targets			
STARD3	0.05	0.00	1.49×10^{-17}
EIF5A	0.08	0.01	4.97×10^{-17}
ERH	0.07	0.01	8.82×10^{-17}
RNA-binding genes			
SUZ12	0.06	0.00	4.49×10^{-18}
C1D	0.11	0.01	8.03×10^{-18}
MRPL23	−0.09	0.01	8.37×10^{-18}
Ribosome genes			
RPL34	−0.07	0.01	8.57×10^{-16}
RPS27	−0.08	0.01	1.35×10^{-15}
RPL39	−0.08	0.01	1.16×10^{-14}
Mitochondrial part genes			
MRPL23	−0.09	0.01	8.37×10^{-18}
MTERF3	−0.06	0.00	3.12×10^{-17}
MICU3	0.09	0.01	3.34×10^{-17}

Meta-analysis results are from all 47 tested tissues, excluding effects from whole blood. The top three most significant genes for each pathway are shown.

NRF1, YY1, GABPB, and E4F1 transcription factors. All of these transcription factors have been implicated in mitochondrial pathways, as ELK1 is associated with the mitochondrial permeability transition pore complex in neurons, NRF1 regulates expression of the mitochondrial translocase TOMM34, YY1 binds to and represses mitochondrial gene expression in skeletal muscle, GABPB is required for mitochondrial biogenesis, and E4F1 controls mitochondrial homeostasis (Barrett et al. 2006; Blesa et al. 2008; Yang et al. 2014; Rodier et al. 2015; Chen et al. 2019). Additionally, we found significant enrichment of signal for genes implicated in ubiquitin-mediated proteolysis and splicing. Given that mitochondrial quality control is regulated through ubiquitination, and that nuclear-encoded spliceosomes are involved in mtRNA splicing, our results likely implicate processes involved in mitochondrial DNA regulatory networks (Bragoszewski et al. 2017; Herai et al. 2017).

mtDNA-CN measured in one tissue has previously been found to be uncorrelated with mtDNA-CN in another tissue from the same individual (Wachsmuth et al. 2016). We found that, although mtRNA transcription in individual tissues was not significantly correlated with blood-derived mtDNA-CN, across all tissues, there was a significant enrichment for positive associations, suggesting a weak positive correlation between blood-derived mtDNA-CN and mtDNA-CN in other tissues. Moreover, we found that blood-derived mtDNA-CN was associated with various biological pathways in nonblood tissues (including mitochondrial function), providing a possible explanation as to why blood-derived mtDNA-CN is associated with aging-related diseases that primarily manifest in nonblood tissues. Further examination of pathways significant in multiple tissues revealed that ribosomal subunit genes were significantly negatively associated with mtDNA-CN. Although there has been conflicting evidence on the relationship between mtDNA-CN and ribosomal content, our study revealed a strong inverse relationship between ribosomal DNA dosage and mtDNA-CN (Gibbons et al. 2014; Guantes et al. 2015). Importantly, because these are statistical associations, causal directionality cannot be determined between gene expression and blood-derived mtDNA-CN. Future follow-up studies are needed to determine functional causality for mtDNA-CN and gene expression.

KEGG pathways that were significantly enriched in multiple tissues included Huntington's disease, Alzheimer's disease, and Parkinson's disease. These aging-related neurodegenerative diseases all have underlying mitochondrial pathologies (Coskun et al. 2010; Petersen et al. 2014; Wei et al. 2017; Park et al. 2018) and dysregulated ubiquitination pathways (Atkin and Paulson 2014). In particular, mtDNA-CN has been implicated in Alzheimer's disease (Rice et al. 2014; Delbarba et al. 2016; Lv et al. 2019) and cognitive function (Lee et al. 2010, 2017a). Further, the ELK1 transcription factor, whose target sequences were significantly enriched in 19 tissues, plays a role in multiple neurodegenerative diseases (Besnard et al. 2011). Finally, after finding that blood-derived mtDNA-CN was associated with expression of neurodegenerative disease genes, we used an independent data set, the UK Biobank, and found that mtDNA-CN was significantly associated with both prevalent and incident neurodegenerative disease risk. In conclusion, our findings show that blood-derived mtDNA-CN is significantly associated with gene expression from tissues across the body and that higher mtDNA-CN is associated with decreased incident neurodegenerative disease risk.

Methods

GTEx sample acquisition

Whole genome sequences were downloaded from the GTEx version 8 cloud repository on 11/18/2020. RNA-sequencing

Table 6. mtDNA-CN is associated with incident and prevalent neurodegenerative disease

Prevalent disease	Odds ratio	Confidence interval	Number of cases/controls	*P*-value
Parkinson's disease	0.90	0.83;0.97	697/368,734	0.005
Dementia (excluding AD)	0.81	0.67;0.99	107/368,607	0.039
Combined neurodegenerative disease	0.89	0.83;0.96	853/368,578	1.00×10^{-3}
Incident disease	**Hazard ratio**	**Confidence interval**	**Number of cases/controls**	***P*-value**
Parkinson's disease	0.92	0.86;0.98	965/367,769	0.016
Alzheimer's disease	0.98	0.91;1.06	705/368,714	0.625
Dementia (excluding AD)	0.93	0.88;0.99	1106/367,501	0.026
Combined neurodegenerative disease	0.95	0.91;0.98	2468/366,110	0.007

Hazard ratios and odds ratios for neurodegenerative disease associate with a 1-standard-deviation increase in whole blood mtDNA-CN estimated from either a Cox proportional-hazards model (incident) or a logistic regression model (prevalent) in the UKBiobank. Analysis was restricted to individuals of European descent, and individuals who were outliers for cell counts were excluded from analysis. *P*-values have not been adjusted for multiple testing.

data used for analyses were downloaded from the GTEx portal (http://gtexportal.org/home/datasets) on 06/18/2019, and phenotypes were obtained from the database of Genotypes and Phenotypes (dbGaP; https://www.ncbi.nlm.nih.gov/gap/) (phs000424.v8.p2).

Estimation of mtDNA-CN

SAMtools version 1.9 (Li et al. 2009) was used to count the number of mitochondrial, unaligned, and total reads for each whole genome sequence. mtDNA-CN was estimated as the number of mitochondrial reads divided by the difference between the number of total reads and the number of unaligned reads to obtain a ratio of mtDNA to nuclear DNA. Whole genome is a highly accurate method for estimation of mtDNA-CN (Wachsmuth et al. 2016; Longchamps et al. 2020).

Correcting mtDNA-CN for covariates

All statistical analyses were performed with R version 3.6.1 (R Core Team 2019). Cell type composition for whole blood samples was determined from RNA sequencing using xCell (Aran et al. 2017), only allowing for deconvolution of cell types found in blood. A stepwise regression in both directions was used to select appropriate cell types to correct mtDNA-CN. To avoid model overfitting, correlated cell types ($R > 0.8$) were removed. The final model used to adjust mtDNA-CN included neutrophils, hematopoietic stem cells, megakaryocytes, subject cohort, ischemic time, age, and sex. Residuals were standardized after adjusting for covariates. Power calculations were performed using R^2 values from previous studies using the pwr package (Longchamps et al. 2020).

Filtering pipeline

A batch effect due to sample collection and/or sequencing methods resulted in significantly altered mtDNA-CN for individuals who were sequenced prior to January 2013. To keep this from confounding the analysis, we excluded subjects with whole genome sequencing prior to January 2013 (Supplemental Fig. S1). Individuals who had greater than 5×10^7 unaligned whole genome sequence reads were also omitted from the analysis. Cell type outliers who were greater than three standard deviations (SDs) from the mean were excluded as well. Only one individual remained from the surgical cohort after filtering and therefore was also removed (Supplemental Fig. S13).

RNA-sequencing pipeline

GTEx version 8 RNA-sequencing data was downloaded from the GTEx website in read counts and normalized using the trimmed mean of M-values method prior to analyses (Robinson and Oshlack 2010; Robinson et al. 2010). For each separate tissue, only genes with expression greater than 0.1 counts for at least 20% of samples for that tissue were retained for analysis. To identify potential hidden confounders, we used surrogate variable analysis (SVA), protecting mtDNA-CN from SV generation (Leek and Storey 2007). SVs were associated with known covariates in the data, such as whether individuals were in the postmortem or the organ donor cohorts (Supplemental Fig. S3). Individuals who were greater than three standard deviations from the mean for the first 10 SVs were omitted from analysis. SV generation was performed iteratively three times.

Linear model for evaluating associations

To reduce the influence of outliers, both the gene expression metric and the mtDNA-CN metric were inverse normal transformed prior to linear regression. We then tested for association using multiple linear regression, with mtDNA-CN as the predictor and gene expression as the outcome, correcting for SVs, sex, cohort, race, ischemic time, and the first three genotyping principal components.

Genomic inflation factor calculation

Genomic inflation factors were calculated by squaring *z*-scores to obtain χ^2 values. The median observed χ^2 value was divided by the expected median to obtain lambda (Devlin and Roeder 1999).

Two-stage permutations

To determine an appropriate *P*-value cutoff, we created null data sets for permutation testing. First, a multiple linear regression model for the alternate hypothesis was used to obtain gene expression residuals. Second, a multiple linear regression model for the null hypothesis was used to obtain estimates for each gene. Residuals from the alternate model were then permuted and added to effect estimates from the null model to create null data sets. Permuted gene expression data were then tested for association with mtDNA-CN. Permutations were performed 1000 times. Minimum *P*-values from each permuted data set were obtained, and the fifth lowest *P*-value was utilized as a permutation cutoff.

Annotation of gene categories

Gene annotations were downloaded from GENCODE (Harrow et al. 2012). Test statistics were then stratified by gene type, and observed and expected distributions were generated for each category.

Overrepresentation of positive beta estimates

Percentage of positive effect estimates was calculated using all nominally significant genes in blood, dividing the number of nominally significant genes with positive effect estimates by the total number of nominally significant genes. Percentages for null distributions were calculated using 1000 permutations, generated using the two-stage permutation method described above.

Gene set enrichment analysis

To examine enrichment for genes in specific pathways, gene sets for KEGG pathways, transcription factor target sequences, and Gene Ontologies were downloaded from the Molecular Signatures Database (Kanehisa and Goto 2000; Liberzon et al. 2011; The Gene Ontology Consortium 2019). Then, using the absolute value of the *t*-scores from the regression model with mtDNA-CN, we performed a *t*-test of *t*-scores for genes in a specific pathway versus genes that were not contained in the pathway. For each tissue, only genes with greater than 0.1 counts in at least 20% of the tissue samples were used for gene set enrichment analyses. Based on which genes passed expression thresholds, different lists of genes were used for enrichment analyses for different tissues. We also performed 1000 permutations using randomized *t*-scores to determine appropriate cutoffs for significance. To confirm that results were not driven by individual genes in a pathway with very large *t*-scores, we also performed *t*-tests using ranked *t*-scores as opposed to absolute value *t*-scores.

REVIGO trimming and visualization of GO terms

For visualization of significantly enriched GO terms and elimination of redundant GO terms, REVIGO (http://revigo.irb.hr/) was

used with the default settings except for the allowed similarity, which was set to medium (0.7) (Supek et al. 2011).

Testing for associations between blood-derived mtDNA-CN and gene expression in other tissues

Filtering parameters and models for testing the association of blood-derived mtDNA-CN with gene expression in other tissues were identical to the pipeline used in whole blood. Only tissues with greater than 50 observations after filtering were tested. For tissues that had no variation in covariates, covariates were dropped from the linear model (i.e., sex was not used in the model for testing gene expression in reproductive organs, and cohort was not used in the model for brain tissues).

Spearman's correlations for effect estimates with whole blood

All significant genes in whole blood that passed the permutation cutoff ($P = 2.38 \times 10^{-6}$) were used for testing. Spearman's correlations between effect estimates in blood and effect estimates in other tissues were calculated. To compare correlations for genes significant in blood with baseline correlation, we randomly selected 100 random genes and calculated correlations between blood estimates and specific tissue estimates for those genes. We repeated this random selection 1000 times to generate multiple baseline correlation measures.

Meta-analysis of genes driving specific ontologies

To calculate meta-analysis effect estimates and *P*-values, the R "*meta*" package (Balduzzi et al. 2019) was used to perform a random-effects meta-analysis using all effect estimates and *P*-values for all tissues, excluding results from whole blood.

Association of mtDNA-CN with neurodegenerative disease in UKB

UKB mtDNA-CN derivation

We started with 49,997 Exome SPB CRAM files (version Jul 2018) downloaded from the UKB data repository and used SAMtools (ver1.9) to extract read summary statistics ("idxstats" command). A custom Perl script was used to aggregate the summary statistics from each individual file into the following categories (see Perl script and example stats file): (1) Total Reads (sum of columns 3 and 4, across all rows); (2) Mapped Reads (sum of column 3, across all rows); (3) Unmapped Reads (sum of column 4 across all rows); (4) Autosomal Reads (sum of column 3, rows 1–22); (5) Chr X; (6) Chr Y; (7) Chr MT; (8) "Random" Reads (sum of column 3, across rows 26–67); (9) "Unknown" Reads (sum of column 3 across rows 68–194); (10) EBV Reads; (11) "Decoy1" Reads (sum of column 3 across rows 196–582); (12) "Decoy2" Reads (sum of column 3 across rows 583–2580). Linear regression models were used to adjust for total DNA and potential technical artifacts. Specifically, we used 10-fold cross-validation for variable selection, using the "leaps" R package (version 3.0), with an initial model with chrMT read count as the dependent variable, and "Total", "Mapped", "unknown", "random", "decoy1", and "decoy2" read counts as the independent variables. For each of the independent variables, we included a natural spline with df = 4 to allow for nonlinear effects. The independent variables "Total", "unknown", "decoy1", and "decoy2" read counts were selected. We then increased the natural spline df to 15 and then used backward selection to reduction model complexity, requiring $P < 0.005$ to keep a term in the model. The final regression model residuals were generated with the following R (version 3.6.0) code:

```
WES.mtDNA = residuals(lm(chrMT~ns(Total, df = 3)
            + ns(unknown, df = 4) + ns(decoy1, df = 7)
            + decoy2))
```

Mitochondrial SNP probe intensities were obtained from the "ukb_chrMT_l2r.txt" file downloaded from the UK Biobank, and samples were stratified by array type (UKBelieve, Axiom). To correct for potential artifacts and/or batch effects, we generated 250 principal components (PCs) using the "rpca" command from the "rsvd" package (version 1.0.3) from autosomal nuclear probes by randomly sampling 5% of probes from either even or odd chromosomes that were required to be present on both array types (n~19,500 probes). Note that we generated the two independent sets of PCs so that we could ensure that probe selection for PCA did not bias results. Prior to PCA, all probe intensities were rank-transformed to reduce the impact of any outliers. For each array type, all mitochondrial SNP probes (UKBelieve, $n = 181$; Axiom, $n = 244$) along with the 250 PCs were regressed on the "WES.mtDNA" metric derived as described above. Beta estimates from these analyses were then used to generate fitted values in the full UK Biobank data set using the "predict" function ("array.mtDNA").

Given the known impact of age, sex, and cell counts on mtDNA-CN, we first used visual inspection to identify outliers for cell counts:

Log(WBC) ≤1.25 or ≥3
Log(RBC) ≤1.4 or ≥2s
Platelet ≤10 or ≥500
Log(Lymphocyte) ≤0.10 or ≥2
Log(Mono) ≥0.9
Log(Neutrophil) ≤0.75 or ≥2.75
Log(Eos) ≥0.75
Log(Baso) ≥0.45

We then excluded self-identified non-white individuals due to insufficient WES data, related individuals (used.in.pca.calculation==0), and cell count outliers and then adjusted for age and sex using the following linear regression model:

```
mtDNA-CN = residuals(lm(array.mtDNA~ns(age, df = 2) + sex
```

Beta estimates from these analyses were then used to generate fitted values in the full UK Biobank data set using the "predict" function.

For all analyses, mtDNA-CN was standardized by subtracting the mean and dividing by the standard deviation.

A Cox proportional-hazards model was used to evaluate the association between mtDNA-CN and time to incident neurodegenerative disease, adjusting for age and sex, whereas a logistic regression model was used to evaluate associations between mtDNA-CN and prevalent neurodegenerative disease. Individuals with prevalent neurodegenerative disease were omitted from the incident analysis.

Software availability

All in-house scripts are available as Supplemental Code and at GitHub (https://github.com/syyang93/mtDNA_GE_scripts).

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by National Institutes of Health grants R01HL13573 and R01HL144569. This research was conducted using data from the Genotype-Tissue Expression (GTEx) project (dbGaP accession: phs000424.v8.p2). The GTEx project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by the National Cancer Institute, the National Human Genome Research Institute, the National Heart, Lung, and Blood Institute, the National Institute on Drug Abuse, the National Institute of Mental Health, and the National Institute of Neurological Disorders and Stroke. This research was also conducted using the UK Biobank Resource under Application Number 17731.

References

Ali AT, Boehme L, Carbajosa G, Seitan VC, Small KS, Hodgkinson A. 2019. Nuclear genetic regulation of the human mitochondrial transcriptome. *eLife* **8:** e41927. doi:10.7554/eLife.41927

Al-Kafaji G, Aljadaan A, Kamal A, Bakhiet M. 2018. Peripheral blood mitochondrial DNA copy number as a novel potential biomarker for diabetic nephropathy in type 2 diabetes patients. *Exp Ther Med* **16:** 1483–1492. doi:10.3892/etm.2018.6319

Aran D, Hu Z, Butte AJ. 2017. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* **18:** 220. doi:10.1186/s13059-017-1349-1

Ashar FN, Moes A, Moore AZ, Grove ML, Chaves PHM, Coresh J, Newman AB, Matteini AM, Bandeen-Roche K, Boerwinkle E, et al. 2015. Association of mitochondrial DNA levels with frailty and all-cause mortality. *J Mol Med* **93:** 177–186. doi:10.1007/s00109-014-1233-3

Ashar FN, Zhang Y, Longchamps RJ, Lane J, Moes A, Grove ML, Mychaleckyj JC, Taylor KD, Coresh J, Rotter JI, et al. 2017. Association of mitochondrial DNA copy number with cardiovascular disease. *JAMA Cardiol* **2:** 1247–1255. doi:10.1001/jamacardio.2017.3683

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. *Nat Genet* **25:** 25–29. doi:10.1038/75556

Atkin G, Paulson H. 2014. Ubiquitin pathways in neurodegenerative disease. *Front Mol Neurosci* **7:** 63. doi:10.3389/fnmol.2014.00063

Balduzzi S, Rücker G, Schwarzer G. 2019. How to perform a meta-analysis with R: a practical tutorial. *Evid Based Mental Health* **22:** 153–160. doi:10.1136/ebmental-2019-300117

Barrett LE, Bockstaele EJV, Sul JY, Takano H, Haydon PG, Eberwine JH. 2006. Elk-1 associates with the mitochondrial permeability transition pore complex in neurons. *PNAS* **103:** 5155–5160. doi:10.1073/pnas.0510477103

Besnard A, Galan-Rodriguez B, Vanhoutte P, Caboche J. 2011. Elk-1 a transcription factor with multiple facets in the brain. *Front Neurosci* **5:** 35. doi:10.3389/fnins.2011.00035

Blesa JR, Prieto-Ruiz JA, Abraham BA, Harrison BL, Hegde AA, Hernández-Yago J. 2008. NRF-1 is the major transcription factor regulating the expression of the human *TOMM34* gene. *Biochem Cell Biol* **86:** 46–56. doi:10.1139/O07-151

Bragoszewski P, Turek M, Chacinska A. 2017. Control of mitochondrial biogenesis and function by the ubiquitin–proteasome system. *Open Biol* **7:** 170007. doi:10.1098/rsob.170007

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562:** 203–209. doi:10.1038/s41586-018-0579-z

Calvo SE, Clauser KR, Mootha VK. 2016. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res* **44:** D1251–D1257. doi:10.1093/nar/gkv1003

Castellani CA, Longchamps RJ, Sumpter JA, Newcomb CE, Lane JA, Grove ML, Bressler J, Brody JA, Floyd JS, Bartz TM, et al. 2020a. Mitochondrial DNA copy number can influence mortality and cardiovascular disease via methylation of nuclear DNA CpGs. *Genomic Med* **12:** 84. doi:10.1186/s13073-020-00778-7

Castellani CA, Longchamps RJ, Sun J, Guallar E, Arking DE. 2020b. Thinking outside the nucleus: mitochondrial DNA copy number in health and disease. *Mitochondrion* **53:** 214–223. doi:10.1016/j.mito.2020.06.004

Chabi B, Mousson de Camaret B, Duborjal H, Issartel J-P, Stepien G. 2003. Quantification of mitochondrial DNA deletion, depletion, and overreplication: application to diagnosis. *Clin Chem* **49:** 1309–1317. doi:10.1373/49.8.1309

Chen H, Sun H, You F, Sun W, Zhou X, Chen L, Yang J, Wang Y, Tang H, Guan Y, et al. 2011. Activation of STAT6 by STING is critical for antiviral innate immunity. *Cell* **147:** 436–446. doi:10.1016/j.cell.2011.09.022

Chen F, Zhou J, Li Y, Zhao Y, Yuan J, Cao Y, Wang L, Zhang Z, Zhang B, Wang CC, et al. 2019. YY1 regulates skeletal muscle regeneration through controlling metabolic reprogramming of satellite cells. *EMBO J* **38:** e99727 doi:10.15252/embj.201899727.

Clay Montier LL, Deng J, Bai Y. 2009. Number matters: control of mammalian mitochondrial DNA copy number. *J Genet Genomics* **36:** 125–131. doi:10.1016/S1673-8527(08)60099-5

Coskun PE, Wyrembak J, Derbereva O, Melkonian G, Doran E, Lott IT, Head E, Cotman CW, Wallace DC. 2010. Systemic mitochondrial dysfunction and the etiology of Alzheimer's disease and down syndrome dementia. *J Alzheimers Dis* **20**(Suppl 2)**:** S293–S310. doi:10.3233/JAD-2010-100351

Delbarba A, Abate G, Prandelli C, Marziano M, Buizza L, Arce Varas N, Novelli A, Cuetos F, Martinez C, Lanni C, et al. 2016. Mitochondrial alterations in peripheral mononuclear blood cells from Alzheimer's disease and mild cognitive impairment patients. *Oxid Med Cell Longev* **2016:** 5923938. doi:10.1155/2016/5923938

Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* **55:** 997–1004. doi:10.1111/j.0006-341X.1999.00997.x

Garcia I, Jones E, Ramos M, Innis-Whitehouse W, Gilkerson R. 2017. The little big genome: the organization of mitochondrial DNA. *Front Biosci (Landmark Ed)* **22:** 710–721. doi:10.2741/4511

The Gene Ontology Consortium. 2019. The Gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Res* **47:** D330–D338. doi:10.1093/nar/gky1055

Gibbons JG, Branco AT, Yu S, Lemos B. 2014. Ribosomal DNA copy number is coupled with gene expression variation and mitochondrial abundance in humans. *Nat Commun* **5:** 4850. doi:10.1038/ncomms5850

Gombart AF, Borregaard N, Koeffler HP. 2005. Human cathelicidin antimicrobial peptide (CAMP) gene is a direct target of the vitamin D receptor and is strongly up-regulated in myeloid cells by 1,25-dihydroxyvitamin D_3. *FASEB J* **19:** 1067–1077. doi:10.1096/fj.04-3284com

The GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45:** 580–585. doi:10.1038/ng.2653

Guantes R, Rastrojo A, Neves R, Lima A, Aguado B, Iborra FJ. 2015. Global variability in gene expression and alternative splicing is modulated by mitochondrial content. *Genome Res* **25:** 633–644. doi:10.1101/gr.178426.114

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774. doi:10.1101/gr.135350.111

Herai RH, Negraes PD, Muotri AR. 2017. Evidence of nuclei-encoded spliceosome mediating splicing of mitochondrial RNA. *Hum Mol Genet* **26:** 2472–2479. doi:10.1093/hmg/ddx142

Herst PM, Rowe MR, Carson GM, Berridge MV. 2017. Functional mitochondria in health and disease. *Front Endocrinol (Lausanne)* **8:** 296. doi:10.3389/fendo.2017.00296

Huang J, Tan L, Shen R, Zhang L, Zuo H, Wang DW. 2016. Decreased peripheral mitochondrial DNA copy number is associated with the risk of heart failure and long-term outcomes. *Medicine (Baltimore)* **95:** e3323. doi:10.1097/MD.0000000000003323

Irizarry RA, Wang C, Zhou Y, Speed TP. 2009. Gene set enrichment analysis made simple. *Stat Methods Med Res* **18:** 565–575. doi:10.1177/0962280209351908

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28:** 27–30. doi:10.1093/nar/28.1.27

Kelly RDW, Mahmud A, McKenzie M, Trounce IA, St John JC. 2012. Mitochondrial DNA copy number is regulated in a tissue specific manner by DNA methylation of the nuclear-encoded DNA polymerase γ A. *Nucleic Acids Res* **40:** 10124–10138. doi:10.1093/nar/gks770

Lee J-W, Park KD, Im J-A, Kim MY, Lee D-C. 2010. Mitochondrial DNA copy number in peripheral blood is associated with cognitive function in apparently healthy elderly women. *Clin Chim Acta* **411:** 592–596. doi:10.1016/j.cca.2010.01.024

Lee J-Y, Kim J-H, Lee D-C. 2017a. Combined impact of telomere length and mitochondrial DNA copy number on cognitive function in community-dwelling very old adults. *DEM* **44:** 232–243. doi:10.1159/000480427

Lee WT, Sun X, Tsai T-S, Johnson JL, Gould JA, Garama DJ, Gough DJ, McKenzie M, Trounce IA, St. John JC. 2017b. Mitochondrial DNA haplotypes induce differential patterns of DNA methylation that result in differential chromosomal gene expression patterns. *Cell Death Discov* **3:** 17062. doi:10.1038/cddiscovery.2017.62

Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3:** 1724–1735. doi:10.1371/journal.pgen.0030161

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009, 1000 Genome Project Data Processing

Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352
Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27:** 1739–1740. doi:10.1093/bioinformatics/btr260
Longchamps RJ, Castellani CA, Yang SY, Newcomb CE, Sumpter JA, Lane J, Grove ML, Guallar E, Pankratz N, Taylor KD, et al. 2020. Evaluation of mitochondrial DNA copy number estimation techniques. *PLoS One* **15:** e0228166. doi:10.1371/journal.pone.0228166
Lv X, Zhou D, Ge B, Chen H, Du Y, Liu S, Ji Y, Sun C, Wang G, Gao Y, et al. 2019. Association of folate metabolites and mitochondrial function in peripheral blood cells in Alzheimer's disease: a matched case-control study. *J Alzheimers Dis* **70:** 1133–1142. doi:10.3233/JAD-190477
Malik AN, Czajka A. 2013. Is mitochondrial DNA content a potential biomarker of mitochondrial dysfunction? *Mitochondrion* **13:** 481–492. doi:10.1016/j.mito.2012.10.011
Márquez-Jurado S, Díaz-Colunga J, das Neves RP, Martinez-Lorente A, Almazán F, Guantes R, Iborra FJ. 2018. Mitochondrial levels determine variability in cell death by modulating apoptotic gene expression. *Nat Commun* **9:** 389. doi:10.1038/s41467-017-02787-4
Meng H, Yaari G, Bolen CR, Avey S, Kleinstein SH. 2019. Gene set meta-analysis with Quantitative Set Analysis for Gene Expression (QuSAGE). *PLoS Comput Biol* **15:** e1006899. doi:10.1371/journal.pcbi.1006899
Mengel-From J, Thinggaard M, Dalgård C, Kyvik KO, Christensen K, Christiansen L. 2014. Mitochondrial DNA copy number in peripheral blood cells declines with age and is associated with general health among elderly. *Hum Genet* **133:** 1149–1159. doi:10.1007/s00439-014-1458-9
Miller FJ, Rosenfeldt FL, Zhang C, Linnane AW, Nagley P. 2003. Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age. *Nucleic Acids Res* **31:** 61e. doi:10.1093/nar/gng060
Moore AZ, Ding J, Tuke MA, Wood AR, Bandinelli S, Frayling TM, Ferrucci L. 2018. Influence of cell distribution and diabetes status on the association between mitochondrial DNA copy number and aging phenotypes in the InCHIANTI study. *Aging Cell* **17:** e12683. doi:10.1111/acel.12683
Nakahira K, Hisata S, Choi AMK. 2015. The roles of mitochondrial damage-associated molecular patterns in diseases. *Antioxid Redox Signal* **23:** 1329–1350. doi:10.1089/ars.2015.6407
Osanai A, Sashinami H, Asano K, Li S-J, Hu D-L, Nakane A. 2011. Mouse peptidoglycan recognition protein PGLYRP-1 plays a role in the host innate immune response against *Listeria monocytogenes* infection. *Infect Immun* **79:** 858–866. doi:10.1128/IAI.00466-10
Park J-S, Davis RL, Sue CM. 2018. Mitochondrial dysfunction in Parkinson's disease: new mechanistic insights and therapeutic perspectives. *Curr Neurol Neurosci Rep* **18:** 21. doi:10.1007/s11910-018-0829-3
Petersen MH, Budtz-Jørgensen E, Sørensen SA, Nielsen JE, Hjermind LE, Vinther-Jensen T, Nielsen SMB, Nørremølle A. 2014. Reduction in mitochondrial DNA copy number in peripheral leukocytes after onset of Huntington's disease. *Mitochondrion* **17:** 14–21. doi:10.1016/j.mito.2014.05.001
R Core Team. 2019. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.
Reddy PH. 2009. The role of mitochondria in neurodegenerative diseases: mitochondria as a therapeutic target in Alzheimer's disease. *CNS Spectr* **14:** 8–18. doi:10.1017/S1092852900024901
Reznik E, Miller ML, Şenbabaoğlu Y, Riaz N, Sarungbam J, Tickoo SK, Al-Ahmadie HA, Lee W, Seshan VE, Hakimi AA, et al. 2016. Mitochondrial DNA copy number variation across human cancers. *eLife* **5:** e10769. doi:10.7554/eLife.10769
Rice AC, Keeney PM, Algarzae NK, Ladd AC, Thomas RR, Bennett JP Jr. 2014. Mitochondrial DNA copy numbers in pyramidal neurons are decreased and mitochondrial biogenesis transcriptome signaling is disrupted in Alzheimer's disease hippocampi. *J Alzheimers Dis* **40:** 319–330. doi:10.3233/JAD-131715
Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11:** R25. doi:10.1186/gb-2010-11-3-r25
Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140. doi:10.1093/bioinformatics/btp616
Rodier G, Kirsh O, Baraibar M, Houlès T, Lacroix M, Delpech H, Hatchi E, Arnould S, Severac D, Dubois E, et al. 2015. The transcription factor E4F1 coordinates CHK1-dependent checkpoint and mitochondrial functions. *Cell Rep* **11:** 220–233. doi:10.1016/j.celrep.2015.03.024
Rusecka J, Kaliszewska M, Bartnik E, Tońska K. 2018. Nuclear genes involved in mitochondrial diseases caused by instability of mitochondrial DNA. *J Appl Genet* **59:** 43–57. doi:10.1007/s13353-017-0424-3
Saha A, Battle A. 2019. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Res* **7:** 1860. doi:10.12688/f1000research.17145.2
Sun X, St John JC. 2018. Modulation of mitochondrial DNA copy number in a model of glioblastoma induces changes to DNA methylation and gene expression of the nuclear genome in tumours. *Epigenetics Chromatin* **11:** 53. doi:10.1186/s13072-018-0223-z
Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS One* **6:** e21800. doi:10.1371/journal.pone.0021800
Tin A, Grams ME, Ashar FN, Lane JA, Rosenberg AZ, Grove ML, Boerwinkle E, Selvin E, Coresh J, Pankratz N, et al. 2016. Association between mitochondrial DNA copy number in peripheral blood and incident CKD in the atherosclerosis risk in communities study. *J Am Soc Nephrol* **27:** 2467–2473. doi:10.1681/ASN.2015060661
Wachsmuth M, Hübner A, Li M, Madea B, Stoneking M. 2016. Age-related and heteroplasmy-related variation in human mtDNA copy number. *PLoS Genet* **12:** e1005939. doi:10.1371/journal.pgen.1005939
Wai T, Ao A, Zhang X, Cyr D, Dufort D, Shoubridge EA. 2010. The role of mitochondrial DNA copy number in mammalian fertility. *Biol Reprod* **83:** 52–62. doi:10.1095/biolreprod.109.080887
Wei W, Keogh MJ, Wilson I, Coxhead J, Ryan S, Rollinson S, Griffin H, Kurzawa-Akinibi M, Santibanez-Koref M, Talbot K, et al. 2017. Mitochondrial DNA point mutations and relative copy number in 1363 disease and control human brains. *Acta Neuropathol Commun* **5:** 13. doi:10.1186/s40478-016-0404-6
West AP, Khoury-Hanold W, Staron M, Tal MC, Pineda CM, Lang SM, Bestwick M, Duguay BA, Raimundo N, MacDuff DA, et al. 2015. Mitochondrial DNA stress primes the antiviral innate immune response. *Nature* **520:** 553–557. doi:10.1038/nature14156
Wood LB, Winslow AR, Strasser SD. 2015. Systems biology of neurodegenerative diseases. *Integr Biol (Camb)* **7:** 758–775. doi:10.1039/C5IB00031A
Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345. doi:10.1038/nature03441
Yang Z-F, Drumea K, Mott S, Wang J, Rosmarin AG. 2014. GABP transcription factor (nuclear respiratory factor 2) is required for mitochondrial biogenesis. *Mol Cell Biol* **34:** 3194–3201. doi:10.1128/MCB.00492-12
Zhang R, Wang Y, Ye K, Picard M, Gu Z. 2017. Independent impacts of aging on mitochondrial DNA quantity and quality in humans. *BMC Genomics* **18:** 890. doi:10.1186/s12864-017-4287-0

Received July 27, 2020; accepted in revised form January 6, 2021.

Blood-derived mitochondrial DNA copy number is associated with gene expression across multiple tissues and is predictive for incident neurodegenerative disease

Stephanie Y. Yang, Christina A. Castellani, Ryan J. Longchamps, et al.

Genome Res. 2021 31: 349-358 originally published online January 13, 2021
Access the most recent version at doi:10.1101/gr.269381.120